PENERAPAN METODE UNDERSAMPLING CLUSTER CENTROIDS PADA ALGORITMA RANDOM FOREST UNTUK PREDIKSI KEBANGKRUTAN PERUSAHAAN

Rangga Wahyu Pratama, Felix John P. Hutabarat, Paskah Abadi Simanullang, Peter Tymoty Hutabarat, Arnita, Fanny Ramadhani
Ilmu Komputer, Universitas Negeri Medan
Jalan William Iskandar Ps. V Medan, Indonesia
Ranggawahyupratama386@mhs.unimed.ac.id

ABSTRAK

Kemajuan teknologi informasi telah mendorong pengembangan model prediksi dalam sektor finansial, termasuk prediksi kebangkrutan perusahaan. Prediksi ini sangat penting bagi pemodal dan investor dalam pengambilan keputusan strategis. Namun, tantangan utama dalam prediksi kebangkrutan adalah ketidakseimbangan kelas, di mana jumlah perusahaan bangkrut jauh lebih sedikit dibandingkan perusahaan yang tidak bangkrut. Penelitian ini bertujuan untuk mengembangkan model klasifikasi yang lebih akurat dalam memprediksi risiko kebangkrutan dengan mengatasi ketidakseimbangan data menggunakan pendekatan yang berbeda dari penelitian sebelumnya. Jika penelitian sebelumnya menggunakan oversampling SMOTE, penelitian ini berkontribusi dengan menerapkan undersampling Cluster Centroid untuk menyeimbangkan distribusi data dari (2834:165) menjadi (165:165). Model dikembangkan menggunakan Random Forest Classifier, dengan evaluasi melalui cross-validation dan hyperparameter tuning untuk meningkatkan performa prediksi. Hasil pengujian menunjukkan bahwa model Random Forest Biasa memiliki akurasi 94,44%, tetapi dengan recall hanya 18,00%, menunjukkan ketidakefektifan dalam mendeteksi perusahaan bangkrut. Setelah menerapkan undersampling Cluster Centroid, akurasi menurun menjadi 63,64%, tetapi recall meningkat signifikan menjadi 89,80%, dengan F1-score 70,97%. Learning curve menunjukkan bahwa model mampu menggeneralisasi data dengan baik tanpa mengalami overfitting yang signifikan. Penelitian ini adalah menunjukkan bahwa Cluster Centroid sebagai metode undersampling dapat meningkatkan kemampuan model dalam mendeteksi kelas minoritas secara lebih efektif dibandingkan oversampling SMOTE, meskipun dengan kompromi pada akurasi keseluruhan. Hasil ini membuka peluang eksplorasi lebih lanjut dalam penggunaan teknik undersampling untuk meningkatkan prediksi kebangkrutan perusahaan.

Kata kunci : *Prediksi Kebangkrutan, Random Forest, Cluster Centroids, Undersampling, Machine Learning*

1. PENDAHULUAN

Tujuan utama didirikannya sebuah perusahaan atau bisnis adalah untuk meningkatkan penjualan, menaikkan nilai saham sehingga mampu menghasilkan keuntungan, serta meningkatkan kesejahteraan para pemegang saham. Dengan perkembangan ekonomi yang terus berubah, banyak perusahaan bersaing secara ketat demi mencapai tingkat kinerja terbaik. Oleh karena itu, pengelolaan keuangan perusahaan menjadi faktor krusial dalam kemajuan bisnis agar tetap bertahan. Dalam upaya mencapai tujuan tersebut, perusahaan akan menghadapi berbagai tantangan, hambatan, serta risiko. Jika tidak mampu mengatasinya, perusahaan berisiko mengalami kebangkrutan [1].

Kebangkrutan menjadi isu krusial dalam aspek keuangan perusahaan. Istilah ini merujuk pada penghentian operasional atau likuidasi perusahaan akibat kegagalan dalam mengelola keuangan, di mana kondisi keuangan tidak sebanding dengan aset yang dimiliki. Dampak dari kebangkrutan ini dapat dirasakan oleh berbagai pihak, termasuk pemegang saham, karyawan, pelanggan, serta investor. Oleh sebab itu, upaya memprediksi kebangkrutan memiliki peran yang signifikan dalam dunia ekonomi dan keuangan [2].

Prediksi merupakan proses pengambilan keputusan terkait operasional bisnis dengan cara yang dapat diandalkan. Prediksi kebangkrutan bertujuan untuk merancang model yang efektif dalam mengidentifikasi risiko kebangkrutan berdasarkan data keuangan perusahaan. Hasil dari prediksi ini dapat dimanfaatkan oleh perusahaan untuk mengevaluasi kondisi keuangan serta merencanakan langkah-langkah pencegahan guna menghindari kemungkinan kebangkrutan [2].

Model prediksi kebangkrutan berbasis *machine learning* telah banyak dikembangkan untuk memperoleh model prediksi yang optimal. Dalam proses klasifikasi, prediksi kebangkrutan termasuk dalam permasalahan ketidakseimbangan kelas *(class imbalance)*, sehingga diperlukan optimasi pada algoritma klasifikasi guna meningkatkan akurasi prediksi. Klasifikasi merupakan proses analisis data yang bertujuan untuk membangun model yang dapat mengidentifikasi serta menjelaskan kategori dalam suatu himpunan data. Algoritma klasifikasi berfungsi untuk memprediksi label kelas dari suatu data, sehingga data tersebut dapat dikategorikan ke dalam

kelas yang telah ditentukan berdasarkan pola yang ditemukan dalam data latih [3].

Dalam penelitian mengenai prediksi kebangkrutan, *Random Forest* merupakan salah satu algoritma yang banyak digunakan karena kemampuannya dalam mengolah data berukuran besar serta menghasilkan klasifikasi yang akurat. Algoritma ini beroperasi dengan membangun sejumlah pohon keputusan dan mengombinasikan hasilnya untuk memperoleh prediksi yang lebih stabil dan andal. Namun, dalam konteks prediksi kebangkrutan, sering muncul permasalahan ketidakseimbangan kelas *(class imbalance)*, di mana jumlah perusahaan yang mengalami kebangkrutan jauh lebih sedikit dibandingkan dengan yang tidak. Kondisi ini dapat menyebabkan model lebih cenderung mengklasifikasikan kelas mayoritas dengan akurasi tinggi, tetapi kurang efektif dalam mendeteksi perusahaan yang berpotensi mengalami kebangkrutan. Oleh karena itu, diperlukan pendekatan optimasi, seperti teknik *Cluster Centroids*, guna meningkatkan kinerja model dalam menangani ketidakseimbangan data tersebut [4], [5].

Selain optimasi data, evaluasi performa model juga menjadi aspek yang sangat penting dalam penelitian ini. Penelitian mengenai prediksi kebangkrutan sudah pernah beberapa kali dilakukan dan menunjukkan bahwa efektivitas model *Random Forest* diukur menggunakan berbagai metrik evaluasi, seperti akurasi, presisi, *recall*, *F1-score*, dan *confusion matrix*, untuk memastikan bahwa model tidak hanya menghasilkan prediksi yang akurat secara keseluruhan, tetapi juga mampu mengenali pola pada kelas minoritas dengan baik [6].

Penelitian sebelumnya oleh Nugroho dan Rilvani menunjukkan bahwa penerapan teknik optimasi, seperti *oversampling* SMOTE, dapat meningkatkan akurasi prediksi hingga 7,40%, yang menegaskan bahwa pemrosesan data sebelum klasifikasi dapat berpengaruh secara signifikan terhadap hasil akhir. Oleh karena itu, penelitian ini tidak hanya berfokus pada penerapan *Random Forest*, tetapi juga membandingkan efektivitas teknik *Cluster Centroids* dalam meningkatkan performa model dalam mengklasifikasikan risiko kebangkrutan perusahaan [4].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis kinerja model Random Forest dalam mengklasifikasikan perusahaan berdasarkan variabel keuangan yang berpengaruh. Selain itu, penelitian ini akan membandingkan performa Random Forest standar dengan Random Forest yang dikombinasikan dengan Cluster Centroids untuk mengatasi ketidakseimbangan kelas dalam prediksi kebangkrutan. Selanjutnya, penelitian ini akan mengevaluasi akurasi model menggunakan metrik seperti akurasi, presisi, recall, F1-score, dan confusion matrix guna menilai kemampuan model dalam mengidentifikasi risiko kebangkrutan. Hasil penelitian ini diharapkan dapat memberikan wawasan mengenai efektivitas metode machine learning dalam analisis kebangkrutan serta mendukung pengembangan model prediksi yang lebih akurat.

2. TINJAUAN PUSTAKA

Prediksi kebangkrutan perusahaan adalah topik yang menarik dalam bidang keuangan dan data mining. Berbagai metode telah dikembangkan untuk meningkatkan akurasi dalam memprediksi kebangkrutan perusahaan. Salah satu metode yang banyak digunakan adalah Random Forest karena kemampuannya dalam mengolah data berukuran besar serta menghasilkan klasifikasi yang akurat [7].

Studi yang dilakukan oleh Nugroho dan Rilvani menunjukkan bahwa penerapan teknik optimasi seperti Synthetic Minority Over-sampling Technique (SMOTE) pada algoritma Random Forest dapat meningkatkan akurasi prediksi kebangkrutan perusahaan hingga 7,40%. Hal ini menegaskan bahwa pemrosesan data sebelum klasifikasi memiliki dampak signifikan terhadap hasil akhir [4].

Selain itu, penelitian yang dilakukan oleh Syamni pada perusahaan batubara menemukan bahwa model prediksi kebangkrutan yang digunakan dapat memberikan informasi penting bagi investor dan pemangku kepentingan untuk mengantisipasi risiko keuangan di masa depan [8].

Algoritma Random Forest telah banyak digunakan dalam berbagai penelitian untuk menangani permasalahan klasifikasi, termasuk prediksi kebangkrutan perusahaan. Penelitian yang dilakukan oleh Bustaman menunjukkan bahwa algoritma Random Forest memiliki keunggulan dalam menangani dataset yang besar serta dapat meningkatkan akurasi prediksi dibandingkan metode tradisional seperti regresi logistik [9].

Selain itu, penelitian oleh Heryanto yang menggunakan model Grover dalam prediksi kebangkrutan menunjukkan bahwa metode ini dapat menjadi alternatif yang efektif. Namun, pendekatan ini tetap membutuhkan optimasi lebih lanjut dalam menangani ketidakseimbangan kelas pada dataset keuangan [10].

Ketidakseimbangan kelas (class imbalance) merupakan salah satu tantangan dalam prediksi kebangkrutan perusahaan. Untuk mengatasi masalah ini, beberapa penelitian telah mengusulkan teknik optimasi seperti undersampling dan oversampling. Penelitian oleh Nugroho dan Rilvani menunjukkan bahwa teknik SMOTE dapat meningkatkan akurasi model Random Forest dalam mendeteksi perusahaan yang berisiko mengalami kebangkrutan. Teknik undersampling Cluster Centroids yang diterapkan dalam penelitian ini juga bertujuan untuk meningkatkan efektivitas model dalam menangani ketidakseimbangan data [4].

Studi oleh Fadli dan Saputra menunjukkan bahwa teknik ini dapat meningkatkan sensitivitas model dalam mengklasifikasikan kelas minoritas [6].

Selain itu, penelitian yang dilakukan oleh Masdiantini dan Warasniasih menyoroti pentingnya analisis laporan keuangan sebagai faktor utama dalam prediksi kebangkrutan [11].

Dengan kombinasi antara teknik machine learning dan analisis laporan keuangan, akurasi prediksi dapat ditingkatkan secara signifikan. Beberapa penelitian telah mengimplementasikan model prediksi kebangkrutan pada berbagai sektor industri. Penelitian oleh Cahya Nugraha menunjukkan bahwa model Grover dapat digunakan dalam menganalisis kebangkrutan perusahaan berdasarkan variabel keuangan tertentu [10].

Sementara itu, penelitian yang dilakukan oleh Septiani dan Wahyuni menggunakan algoritma K-Means Clustering untuk mengelompokkan tingkat kepuasan pelanggan dalam layanan bisnis [5].

Meskipun penelitian ini tidak secara langsung terkait dengan prediksi kebangkrutan, pendekatan yang digunakan dapat diadaptasi untuk menganalisis pola kebangkrutan berdasarkan karakteristik perusahaan. Studi yang dilakukan oleh Putu Riesty Masdiantini dan Ni Made Sindy Warasniasih juga menunjukkan bahwa laporan keuangan memiliki peran penting dalam memprediksi kebangkrutan perusahaan. Dengan menganalisis rasio keuangan tertentu, risiko kebangkrutan dapat diidentifikasi lebih dini [11].

3. METODE PENELITIAN

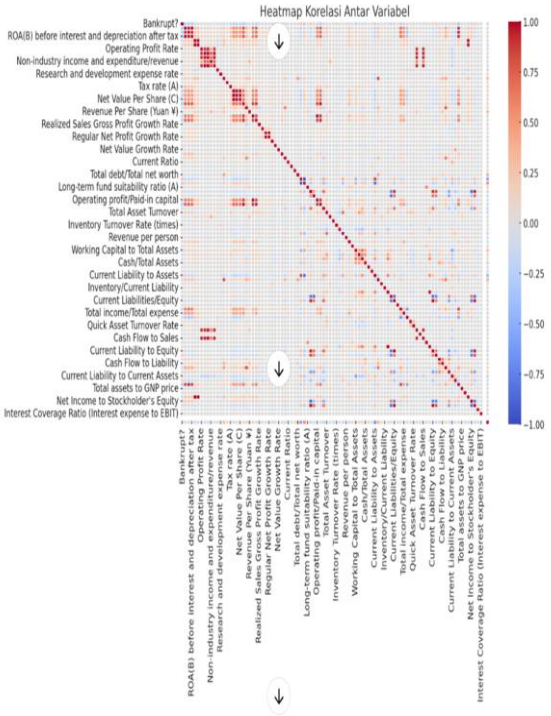
Penelitian ini menggunakan metode eksperimen dengan pendekatan machine learning untuk melakukan prediksi kebangkrutan perusahaan. Algoritma utama yang digunakan dalam penelitian ini adalah Random Forest, yang dikombinasikan dengan teknik undersampling Cluster Centroids untuk menangani masalah ketidakseimbangan kelas pada dataset.

4. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle dan berisi informasi keuangan perusahaan yang digunakan untuk memprediksi kebangkrutan. Dataset ini terdiri dari 2.999 record dengan 96 fitur, seluruhnya dalam bentuk numerik. Karena tujuan penelitian ini adalah klasifikasi, algoritma yang digunakan adalah Random Forest, yang dikenal memiliki kemampuan menangani dataset dengan banyak fitur dan memberikan interpretasi penting terhadap variabel yang berkontribusi dalam prediksi. Untuk meningkatkan performa model, penelitian ini juga menerapkan teknik undersampling Cluster Centroids guna mengatasi ketidakseimbangan kelas. Analisis terhadap dataset ini dilakukan melalui beberapa tahap, mulai dari eksplorasi data hingga evaluasi model.

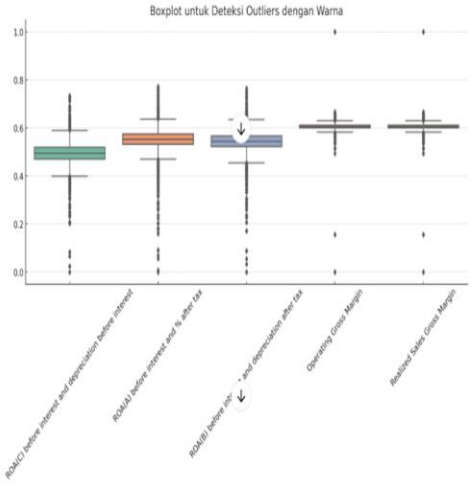
4.1. Hasil Eksplorasi Data (EDA)

Dalam tahap awal eksplorasi data, dilakukan pengecekan terhadap keberadaan missing values dan outliers. Missing values dapat mengganggu proses analisis dan model prediksi, sementara outliers dapat menyebabkan bias pada hasil model. Oleh karena itu, deteksi dan penanganan keduanya menjadi langkah penting sebelum melanjutkan ke tahap pemodelan. Pemeriksaan missing values menunjukkan bahwa dataset tidak memiliki nilai yang hilang. Selanjutnya, deteksi outliers dilakukan menggunakan metode Z-Score dengan threshold > 3 untuk mengidentifikasi data yang berada jauh dari distribusi normal. Dari hasil analisis, ditemukan bahwa beberapa fitur memiliki jumlah outlier yang cukup tinggi, seperti ROA(A), ROA(B), dan Operating Gross Margin. Nilai-nilai ekstrem ini berpotensi mempengaruhi performa model prediksi. Jika tidak ditangani, model dapat mengalami kesulitan dalam menggeneralisasi pola data dengan baik.



Gambar 1. Heatmap korelasi antar fitur

Heatmap korelasi menunjukkan adanya beberapa fitur yang memiliki korelasi tinggi satu sama lain. Misalnya, fitur Current Liability to Liability dan Total Debt to Total Assets menunjukkan korelasi yang kuat, yang dapat mengindikasikan adanya informasi yang tumpang tindih. Fitur yang memiliki korelasi tinggi berpotensi menyebabkan multikolinearitas, yang dapat mengurangi akurasi model prediksi dan meningkatkan risiko overfitting. Oleh karena itu, perlu dilakukan seleksi fitur untuk mengurangi dampak korelasi yang berlebihan. Analisis distribusi data bertujuan untuk memahami karakteristik masing-masing fitur, termasuk keberadaan skewness, penyebaran nilai, dan kemungkinan anomali.



Gambar 2. Boxplot deteksi outliers

Boxplot menunjukkan bahwa beberapa fitur memiliki distribusi yang sangat tidak merata, dengan outlier yang signifikan. Fitur seperti ROA(A), ROA(B), dan Operating Gross Margin memiliki banyak titik ekstrem di luar rentang interkuartil. Hal ini mengindikasikan bahwa dataset memiliki distribusi yang condong (skewed), yang dapat menyebabkan model lebih condong ke nilai tertentu. Jika diperlukan, metode transformasi seperti log transformation atau winsorization dapat digunakan untuk mengurangi dampak outlier. Multikolinearitas terjadi ketika beberapa variabel independen memiliki hubungan yang sangat kuat, yang dapat menyebabkan ketidakstabilan dalam estimasi model. Untuk mengevaluasi multikolinearitas, digunakan Variance Inflation Factor (VIF).

Gambar 3. Distribusi kelas data

Hasil perhitungan VIF menunjukkan bahwa beberapa fitur mengalami multikolinearitas tinggi. Misalnya, fitur Current Liability to Liability memiliki VIF yang sangat besar, menunjukkan bahwa fitur tersebut dapat memberikan informasi yang berlebihan terhadap model. Secara umum, fitur dengan VIF > 10 sebaiknya dikurangi atau dilakukan seleksi fitur agar model lebih stabil dan dapat menghindari overfitting. Distribusi kelas dalam dataset sangat penting untuk memahami keseimbangan data antara kategori target (perusahaan bangkrut vs tidak bangkrut). Ketidakseimbangan kelas yang besar dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas, sehingga menurunkan akurasi dalam mengklasifikasikan kelas minoritas.

4.2. Feature Selection

Dalam analisis data, tidak semua fitur memiliki kontribusi yang signifikan terhadap hasil prediksi. Beberapa fitur mungkin hanya menambah kompleksitas tanpa meningkatkan performa model, sementara yang lain bisa memiliki korelasi tinggi dengan target tetapi menyimpan informasi yang redundan. Oleh karena itu, diperlukan proses Feature Selection untuk memilih fitur yang benar-benar relevan dan memiliki dampak besar dalam menentukan apakah suatu perusahaan akan mengalami kebangkrutan atau tidak. Feature selection dalam penelitian ini dilakukan menggunakan dua pendekatan utama, yaitu Embedded Feature Selection dengan Random Forest Importance dan Filter-Based Selection menggunakan Mutual Information atau Korelasi Pearson. Pada metode Embedded Feature Selection, digunakan Random Forest Importance untuk menilai kontribusi setiap fitur dalam membedakan perusahaan yang bangkrut dan tidak bangkrut.

Random Forest bekerja dengan membangun banyak pohon keputusan secara acak, di mana setiap pohon melakukan pemisahan data berdasarkan berbagai fitur. Setelah pohon-pohon ini terbentuk, algoritma mengukur seberapa sering fitur tertentu digunakan dalam membagi data dengan efektif. Semakin sering sebuah fitur berperan dalam pemisahan yang meningkatkan akurasi model, semakin tinggi bobot kepentingannya.

Dengan cara ini, fitur-fitur yang kurang relevan atau memiliki kontribusi kecil dalam prediksi kebangkrutan dapat dieliminasi atau dikesampingkan. Sementara itu, Filter-Based Selection menggunakan dua teknik berbeda: Mutual Information dan Korelasi Pearson. Mutual Information mengukur hubungan informasi antara fitur dan target tanpa mengasumsikan hubungan linear, sehingga dapat menangkap pola yang kompleks dalam data. Jika Mutual Information antara fitur dan target tinggi, berarti fitur tersebut memberikan banyak informasi yang berguna dalam menentukan kebangkrutan perusahaan. Di sisi lain, Korelasi Pearson mengukur hubungan linear antara fitur dan target, dengan nilai korelasi berkisar antara -1 hingga 1. Jika korelasi terlalu rendah, fitur tersebut mungkin tidak terlalu berguna dalam prediksi; sebaliknya, jika korelasi sangat tinggi, fitur tersebut bisa memiliki informasi yang redundan dengan fitur lain.

Dari hasil Feature Selection, didapat 10 fitur utama yang memiliki pengaruh paling besar dalam menentukan kemungkinan kebangkrutan perusahaan. Tabel berikut menampilkan fitur-fitur tersebut, lengkap dengan deskripsi singkat agar mudah

dipahami, bahkan oleh mereka yang tidak memiliki latar belakang keuangan. Selain itu, fitur-fitur ini juga dikategorikan berdasarkan jenis metrik finansialnya, seperti profitabilitas, struktur modal, dan biaya keuangan, untuk memberikan gambaran lebih jelas mengenai peran setiap fitur dalam analisis kebangkrutan.

Tabel 1. Hasil seleksi fitur

No	Nama Fitur	Kategori Finansial	Keterangan
1	Per Share Net profit before tax (Yuan ¥)	Profitabilitas	Laba sebelum pajak per lembar saham (menunjukkan profitabilitas per saham)
2	Persistent EPS in the Last Four Seasons	Profitabilitas	Laba per saham (EPS) selama 4 musim terakhir, menunjukkan konsistensi keuntungan.
3	Net Income to Stockholder's Equity	Profitabilitas	Rasio laba bersih terhadap ekuitas pemegang saham (mengukur profitabilitas investasi pemegang saham).
4	Net profit before tax/Paid-in capital	Profitabilitas	Perbandingan laba sebelum pajak terhadap modal yang disetor, menunjukkan efisiensi penggunaan modal.
5	Borrowing dependency	Struktur Modal	Ketergantungan perusahaan terhadap pinjaman untuk operasi bisnis.
6	Continuous interest rate (after tax)	Biaya Keuangan	Suku bunga efektif setelah pajak yang diterapkan secara berkelanjutan.
7	Total debt/Total net worth	Struktur Modal	Rasio total utang terhadap kekayaan bersih, menunjukkan leverage keuangan perusahaan.
8	Debt ratio %	Struktur Modal	Persentase total utang dibandingkan total aset, semakin tinggi berarti semakin berisiko.
9	Liability to Equity	Struktur Modal	Rasio kewajiban terhadap ekuitas, mengukur seberapa besar aset dibiayai oleh utang dibanding ekuitas.
10	Net Income to Total Assets	Profitabilitas	Rasio laba bersih terhadap total aset, mengukur efisiensi penggunaan aset dalam menghasilkan laba.

4.3. Menyeimbangkan Data (Handling Imbalanced Data)

Setelah melakukan pemilihan fitur terbaik menggunakan 10 fitur yang paling relevan, kami melanjutkan ke tahap analisis distribusi label pada dataset. Dari hasil analisis, diketahui bahwa distribusi label pada dataset awal sangat tidak seimbang, dengan jumlah data untuk label 0 (tidak bangkrut) mencapai 2834 , sementara jumlah data untuk label 1 (bangkrut) hanya 165 . Ketidakseimbangan ini dapat menyebabkan model machine learning cenderung bias terhadap kelas mayoritas, sehingga performa prediksi untuk kelas minoritas menjadi kurang optimal. Untuk mengatasi masalah ini, kami menerapkan teknik undersampling menggunakan Cluster Centroids. Cluster Centroids adalah metode undersampling yang bekerja dengan mengelompokkan data mayoritas ke dalam cluster-cluster, kemudian memilih satu representasi (centroid) dari setiap cluster untuk menggantikan data aslinya. Dengan cara ini, jumlah data mayoritas dapat dikurangi secara signifikan tanpa kehilangan informasi penting yang terkandung dalam pola datanya. Setelah proses undersampling menggunakan Cluster Centroids , distribusi label menjadi seimbang, dengan masing-masing kelas memiliki 165 data . Dengan dataset yang seimbang, model diharapkan dapat belajar secara lebih adil dan memberikan hasil prediksi yang lebih akurat untuk kedua kelas.

Tabel 2. Data train dengan cluster centroid

Label (y)	Row data	Row data with cluster centroid
0	2834	165
1	165	165

Setelah dataset diseimbangkan menggunakan Cluster Centroids, tahap selanjutnya adalah langkah 4: Perancangan & Training Model . Pada tahap ini, kami menggunakan Random Forest Classifier sebagai baseline model untuk membangun prediksi awal. Dataset dibagi menjadi 60% data latih dan 40% data uji , dengan evaluasi awal dilakukan tanpa tuning parameter untuk melihat performa dasar model dalam memprediksi label 0 dan 1. Pendekatan ini bertujuan untuk mengevaluasi sejauh mana model dapat memanfaatkan dataset yang telah seimbang.

4.4. Perancangan & Pengujian Model

Setelah melakukan fitur seleksi yang menghasilkan 10 fitur terbaik dan mengaplikasikan undersampling Cluster Centroid untuk menangani ketidakseimbangan kelas pada dataset, data yang telah dire-sample kini menjadi lebih seimbang, dengan jumlah sampel yang serupa antara kelas minoritas dan mayoritas. Data yang telah seimbang ini siap untuk diproses lebih lanjut. Selanjutnya, pembagian data dilakukan dengan menggunakan rasio 70% untuk data latih dan 30% untuk data uji. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji performa model pada data yang belum pernah dilihat sebelumnya. Pembagian ini bertujuan untuk memastikan bahwa model tidak hanya bekerja

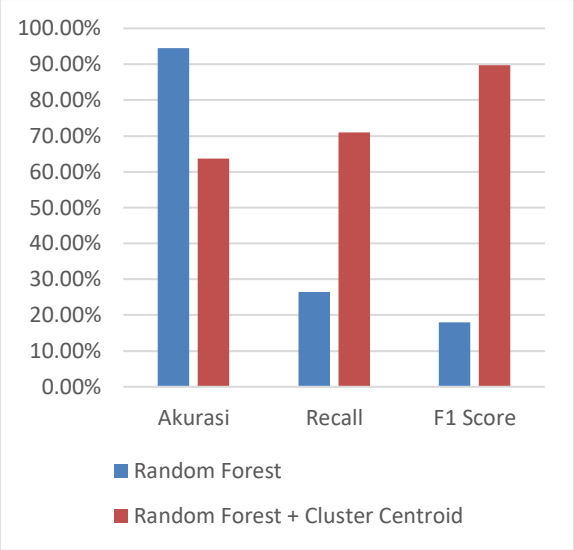
dengan baik pada data yang dilatih, tetapi juga pada data yang baru, yang menyerupai kondisi dunia nyata. Tahap berikutnya adalah prediksi menggunakan algoritma Random Forest Classifier, dengan pengujian menggunakan cross-validation untuk menilai kinerja model. Pada tahap ini, dilakukan dua kali pengujian untuk algoritma Random Forest Classifier, yaitu pertama kali menggunakan data awal sebelum dilakukan re-sampling dan kedua kali menggunakan data yang sudah dilakukan re-sampling dengan Cluster Centroid. Tujuan dari pengujian ini adalah untuk melihat perbandingan hasil performa model sebelum dan sesudah penerapan teknik re-sampling untuk mengatasi ketidakseimbangan kelas.

Pengujian pertama dilakukan dengan menggunakan dataset awal, yang memiliki distribusi kelas yang sangat tidak seimbang. Dalam pengujian ini, algoritma Random Forest Classifier diuji menggunakan cross-validation, menghasilkan akurasi sebesar 94,44%. Meskipun akurasi yang cukup tinggi tercapai, metrik lainnya seperti recall dan F1-score menunjukkan kekurangan model dalam mendeteksi kelas minoritas. Nilai recall yang diperoleh adalah 18,00%, dan F1-score hanya mencapai 26,47%, yang menunjukkan bahwa model memiliki kesulitan dalam memprediksi kelas minoritas dengan tepat. Pada pengujian kedua, dilakukan undersampling dengan Cluster Centroid untuk menyeimbangkan distribusi kelas. Data hasil re-sampling ini kemudian digunakan untuk melatih model dengan algoritma yang sama, yaitu Random Forest Classifier dengan cross-validation. Hasil pengujian kedua menunjukkan penurunan akurasi menjadi 63,64%, tetapi ada peningkatan signifikan dalam metrik lainnya. Nilai recall meningkat menjadi 89,80%, dan F1-score menjadi 70,97%, yang menunjukkan bahwa meskipun akurasi menurun, model lebih mampu mendeteksi kelas minoritas setelah re-sampling.

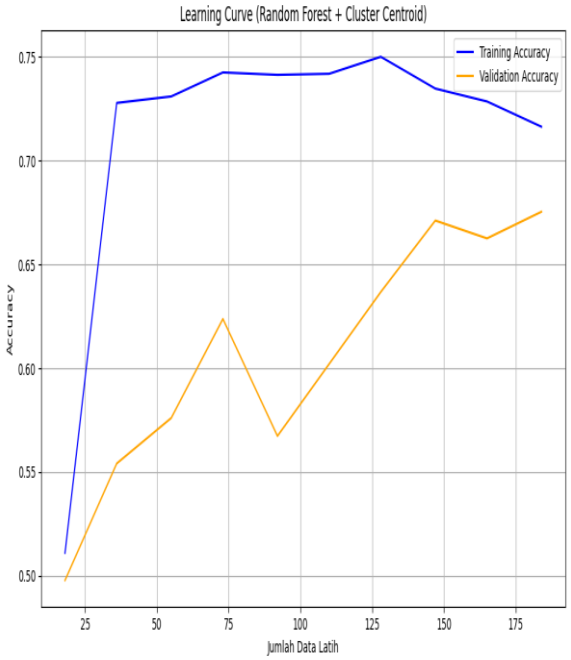
Tabel 3. Hasil pengujian

Model	Akurasi	Recall	F1 Score
Random Forest	94.44%	26.47%	18.00%
Random Forest + Cluster Centroid	63.64%	70.97%	89.80%

Dari hasil ini, dapat disimpulkan bahwa re-sampling menggunakan Cluster Centroid berhasil meningkatkan kemampuan model dalam mendeteksi kelas minoritas, meskipun akurasi keseluruhan mengalami penurunan. Teknik ini terbukti efektif dalam meningkatkan recall dan F1-score, yang menunjukkan bahwa model lebih baik dalam mendeteksi perusahaan yang bangkrut (kelas minoritas). Namun, meskipun ada peningkatan, masih diperlukan optimasi lebih lanjut untuk menyeimbangkan akurasi dan F1-score agar model dapat lebih stabil.



Gambar 4. Grafik perbandingan akurasi



Gambar 5. Learning curve

Performa hasil training dataset menggunakan algoritma Random Forest Classifier dengan undersampling Cluster Centroid disajikan pada grafik berikut. Grafik tersebut menjelaskan bahwa model mengalami overfitting pada tahap awal pelatihan, yang dapat dilihat dari perbedaan yang cukup besar antara performa training accuracy dan validation accuracy. Pada titik tertinggi, training accuracy mencapai sekitar 0.75, sedangkan validation accuracy berada pada angka yang lebih rendah, yaitu sekitar 0.68. Perbedaan ini menunjukkan bahwa model belajar dengan sangat baik pada data latih, namun kurang mampu dalam mengeneralisasi data uji (validation data), yang mengindikasikan adanya overfitting. Namun, setelah beberapa iterasi, performa training accuracy mulai stabil di sekitar 0.72, sementara validation accuracy

juga menunjukkan sedikit peningkatan, mencapai sekitar 0.68 pada titik terakhir. Meskipun kedua kurva mulai mendekat dan menunjukkan tren yang lebih datar, masih terdapat sedikit gap antara keduanya. Ini menunjukkan bahwa meskipun model sudah cukup baik dalam memprediksi data uji, masih ada ruang untuk perbaikan dalam hal generalisasi agar validation accuracy bisa lebih mendekati training accuracy, yang akan menunjukkan bahwa model

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian model, dapat disimpulkan bahwa penggunaan Random Forest + Cluster Centroid memberikan perubahan signifikan dalam kinerja model dibandingkan dengan Random Forest Biasa. Pada model Random Forest Biasa, akurasi mencapai 94,44%, namun recall hanya 18,00% dan F1-score 26,47%, yang menunjukkan kesulitan model dalam mendeteksi kelas minoritas. Sementara itu, setelah penerapan undersampling Cluster Centroid, meskipun akurasi menurun menjadi 63,64%, recall meningkat drastis menjadi 89,80%, dan F1-score menjadi 70,97%, menunjukkan kemampuan model yang lebih baik dalam mendeteksi kelas minoritas. Dalam learning curve yang dihasilkan, terlihat bahwa meskipun ada perbedaan antara training accuracy dan validation accuracy pada awalnya, kedua kurva mulai mendekat dan stabil seiring waktu. Pada awalnya, training accuracy lebih tinggi, mengindikasikan potensi overfitting. Namun, seiring iterasi, perbedaan keduanya menyusut, menunjukkan bahwa model mulai mampu menggeneralisasi lebih baik dan tidak mengalami overfitting signifikan. Meskipun masih ada sedikit gap, model menunjukkan perbaikan dalam memprediksi data baru.

DAFTAR PUSTAKA

[1] F. F. Rozi and Damayanti, "Analisis Kebangkrutan Melalui Perbandingan antara Model Altman Z-Scoredan Springate pada Perusahaan Industri Makanan dan Minuman Yang Terdaftar Di Bursa Efek Indonesia," *Bisman (Bisnis dan Manajemen): The Journal Of Business and Management*, vol. 5, no. 1, pp. 46–58, 2022, [Online]. Available: www.idx.com.

[2] S. D. Saladi and R. Yarlagadda, "An enhanced bankruptcy prediction model using fuzzy clustering model and random forest algorithm," *Revue d'Intelligence Artificielle*, vol. 35, no. 1, pp. 77–83, Feb. 2021, doi: 10.18280/ria.350109.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. 2012.

[4] A. Nugroho and E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan Application of the SMOTE Oversampling Method to the Random Forest Algorithm for Predicting Company Bankruptcy," *Techno.COM*, vol. 22, no. 1, pp. 207–214, 2023.

[5] N. Septiani and S. Wahyuni, "Implementasi Data Mining Dalam Mengelompokkan Tingkat Kepuasan Pemakaian Jasa Cleaning Service Dengan Menggunakan Algoritma K-Means Clustering," *Bulletin of Information Technology (BIT)*, vol. 5, no. 4, pp. 340–354, 2024, doi: 10.47065/bit.v5i2.1729.

[6] M. Fadli and R. A. Saputra, "KLASIFIKASI DAN EVALUASI PERFORMA MODEL RANDOM FOREST UNTUK PREDIKSI STROKE Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," *JT : Jurnal Teknik*, vol. 12, no. 02, pp. 72–80, Oct. 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index

[7] R. Rustam, A. Bustaman, and D. Sarwinda, "Prediksi Kebangkrutan Bank dengan Menggunakan Random Forest," *Jurnal Teknik Informatika*, vol. 9, no. 1, pp. 120–135, 2025.

[8] G. Syamni, M. S. A. Majid, and H. Siregar, "Analisis Akurasi Model-Model Pemprediksi Kebangkrutan: Studi pada Perusahaan Batubara," *Jurnal Ekonomi dan Keuangan*, vol. 15, no. 2, pp. 87–99, 2023.

[9] M. Bustaman, R. Rustam, and A. Sarwinda, "Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM," *Jurnal Sistem dan Informatika Bisnis*, vol. 8, no. 3, pp. 45–56, 2024.

[10] A. C. N. Heryanto, "Analisis Prediksi Kebangkrutan Perusahaan dengan Model Grover," *Jurnal Manajemen dan Keuangan*, vol. 10, no. 1, pp. 55–67, 2023.

[11] P. R. Masdiantini and N. M. S. Warasniasih, "Laporan Keuangan dan Prediksi Kebangkrutan Perusahaan," *Jurnal Akuntansi dan Keuangan*, vol. 18, no. 1, pp. 33–45, 2024.